Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Information Theoretical Aspects of Complex Systems

Lecture 2.03

EEU45C09 / EEP55C09
Self Organising Technological Networks

CONNECT
Networks of the Future

Nicola Marchetti
nicola.marchetti@tcd.ie

Encoding efficiency vs. entropy

- In building encoding schemes, we have to use our best understandings of the *structure* of a data stream (in other words, we want to use our best *probability model* of the data stream)

- The *entropy* gives us a lower bound on our encoding efficiency. Thus, if we want to improve our schemes, we will have to develop successively better probability models

Scientific theories vs. entropy

- One way to think about a scientific theory is that a theory is just an efficient way of encoding (i.e., *structuring*) our knowledge about (some aspect of) the world.

- A *good theory* is one which reduces the (relative) entropy of our (probabilistic) understanding of the system (i.e., that decreases our average *lack of knowledge* about the system)

Noisy channels

- Shannon went on to generalise to the (more realistic) situation in which the channel is *noisy*

- In other words, not only are we unsure about the data stream we will be transmitting (encoded) through the channel, but the channel itself adds an additional layer of *uncertainty/probability* to our transmissions

- Given a source of symbols and a channel with noise (in particular, given probability models for the source and the channel noise), we can talk about the *capacity of the channel*

- We work with two sets of symbols, the input symbols and the output symbols

Conditional probability

- Given two RVs X, Y, taking values in A, B we denote their joint probability as $p_{X,Y}(x, y)$
- The conditional probability for Y given X is indicated by $p_{Y|X}(y|x)$ and we can calculate it as

$$p(y|x) = \frac{p(x, y)}{p(x)}$$

- When the RVs X, Y are independent, $p(y|x)$ is x-independent, i.e. $p(y|x)=p(y)$

Noisy channels

- Let us say the two sets of symbols are $A=\{a_1,a_2,...,a_n\}$ and $B=\{b_1,b_2,...,b_m\}$. Note that we do not necessarily assume the same number of symbols in the two sets

- Given the noise in the channel, when symbol b_j comes out of the channel, we cannot be certain which a_i was put in. The channel is characterized by the set of probabilities $\{P(a_i|b_j)\}$

- We can then consider various related information and entropy measures

Mutual Information

- First, we can consider the information we get from observing a symbol b_j

- Given a probability model of the source, we have an *a priori* estimate $P(a_i)$ that symbol a_i will be sent next

- Upon observing b_j we can revise our estimate to $P(a_i|b_j)$

- The change in our (*mutual*) information is

$$I(a_i;b_j) = \log\left(\frac{1}{P(a_i)}\right) - \log\left(\frac{1}{P(a_i|b_j)}\right) = \log\left(\frac{P(a_i|b_j)}{P(a_i)}\right)$$

Mutual Information - Properties

- We have the properties

  ✓ $I(a_i; b_j) \leq I(a_i)$

  ✓ $I(a_i; b_j) = I(a_i) - I(a_i | b_j)$

  ✓ $I(a_i; b_j) = I(b_j; a_i)$

Use Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- If a_i and b_j are independent (i.e., if $P(a_i, b_j) = P(a_i) \cdot P(b_j)$) then $I(a_i; b_j) = 0$

- Averaging the mutual information over all the symbols:

$$I(A; b_j) = \sum_i P(a_i | b_j) \cdot I(a_i; b_j) = \sum_i P(a_i | b_j) \cdot \log\left(\frac{P(a_i | b_j)}{P(a_i)}\right)$$

Mutual Information – Properties

- Thus

$$
\begin{aligned}
I(A;B) &= \sum_j P(b_j) \cdot I(A;b_j) = \\
&= \sum_j P(b_j) \cdot \sum_i P(a_i|b_j) \log\left(\frac{P(a_i|b_j)}{P(a_i)}\right) \\
&= \sum_j \sum_i P(a_i,b_j) \log\left(\frac{P(a_i,b_j)}{P(a_i)P(b_j)}\right) \\
&= I(B;A)
\end{aligned}
$$

- $I(A;B) \geq 0$

- $I(A;B)=0$ if and only if A,B independent

Sequences of RVs and Markov Chains

- A random process generates a *sequence of Random Variables (RV)* $\{X_t\}_{t \in \aleph}$, each taking values in some space A

- We denote by $P_N(x_1, \ldots, x_N)$ the joint probability distribution of the first *N* variables

- The sequence $\{X_t\}_{t \in \aleph}$ is said to be a *Markov chain* if

$$P_N(x_1, \ldots, x_N) = p_1(x_1) \prod_{t=1}^{N-1} w(x_t \to x_{t+1})$$

Sequences of RVs and Markov Chains (2)

- $\{p_1(x)\}_{x\in A}$ is called the initial state, and

$\{w(x \to y)\}_{x,y\in A}$ are the *transition* probabilities of the chain

- The transition probabilities must be (nonnegative and) *normalised*

$$\sum_{y\in A} w(x \to y) = 1$$

Data Processing Inequality

- The mutual information gets *degraded* when data is transmitted or processed

- This fact is quantified by the so-called *data processing inequality*

- **Proposition.**

  - ✓ Consider a Markov chain $X \to Y \to Z$ (so that the joint probability of the three RVs can be written as $p_1(x) w_2(x \to y) w_3(y \to z)$). Then

    - ❖ $I_{X,Z} \leq I_{X,Y}$

    - ❖ If $Z=f(Y)$ we have that $I_{X,f(Y)} \leq I_{X,Y}$ (in other words, f degrades the information)

Entropy

- The *entropy* S_X of discrete RV X with probability density $p(x)$ is defined as

$$S_X \equiv -\sum_{x \in A} p(x)\log(p(x)) = \mathrm{E}[\log(1/p(X))]$$

where A is the set of values X can take

- The entropy gives a measure of the *uncertainty* of the RV

Entropy - Properties

- $S_X \geq 0$

- $S_X = 0$ if and only if the RV X is *certain* ➔ X takes one value with probability one

- Among all probability distributions on a set A with M elements, S_X is maximum when all events x are equiprobable, with $p(x)=1/M$. The entropy is then $S_X=log(M)$

- If X,Y are two independent RV (meaning that $p_{X,Y}(x,y) = p_X(x)p_Y(y)$) then

$$S_{X,Y} = -\sum_{x,y} p_{X,Y}(x,y)\log\left[p_{X,Y}(x,y)\right] = S_X + S_Y$$

Try to prove this.

- $S_{X,Y} \leq S_X + S_Y$ (generalisable to n RV's)

Entropy Rate

- When we have a sequence of RVs generated by a random process, it is intuitively clear that the entropy grows with the number N of variables

- This intuition suggests to define the *entropy rate* of a sequence $\{X_t\}_{t\in\aleph}$ as

$$s_X = \lim_{N\to\infty} \frac{S_{X_1,\dots,X_N}}{N} =$$
$$= -\lim_{N\to\infty} \frac{\sum_{x_1,\dots,x_N} p_{X_1,\dots,X_N}(x_1,\dots,x_N)\log\left[p_{X_1,\dots,X_N}(x_1,\dots,x_N)\right]}{N}$$

Conditional Entropy

- When X, Y are dependent, it is interesting to have a measure on their degree of dependence

  - ✓ How much information does one obtain on the value of y if one knows x ?

  - ✓ The notions of conditional entropy and mutual information will be useful in this respect

- Let us define the *conditional entropy* $H_{Y|X}$ as the entropy of the distribution $p(y|x)$, averaged over x

$$S_{Y|X} \equiv -\sum_{x \in A} p(x) \sum_{y \in B} p(y|x) \log[p(y|x)]$$

Conditional Entropy and Mutual Entropy

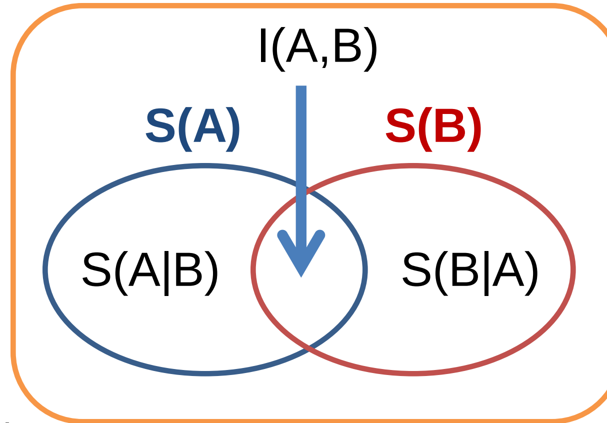
$$S(A) = \sum_{i=1}^{n} P(a_i) \cdot \log\left(1/P(a_i)\right)$$

$$S(B) = \sum_{j=1}^{m} P(b_j) \cdot \log\left(1/P(b_j)\right)$$

$$S(A/B) = \sum_{j=1}^{m} P(b_j) \sum_{i=1}^{n} P(a_i \mid b_j) \cdot \log\left(1/P(a_i \mid b_j)\right)$$

$$S(A, B) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(a_i, b_j) \cdot \log\left(1/P(a_i, b_j)\right)$$

Mutual Information and Entropy



$$S(A,B) = S(A) + S(B|A)$$
$$= S(B) + S(A|B)$$

- And this is how mutual information is related to mutual entropy

$$I(A;B) = S(A) + S(B) - S(A,B)$$
$$= S(A) - S(A|B)$$
$$= S(B) - S(B|A)$$
$$\geq 0$$

I(A;B) = 0 only when *A*,*B* are independent as in that case S*(A,B)*=S*(A)*+S*(B)*

- The mutual information measures the information that *A* and *B* *share*: it measures how much knowing one of these variables reduces uncertainty about the other, while mutual entropy measures the *total* information we get out of *A* and *B*

Acknowledgment

- The material for this lecture has been inspired by

http://www.stanford.edu/~montanar/RESEARCH/BOOK/partA.pdf